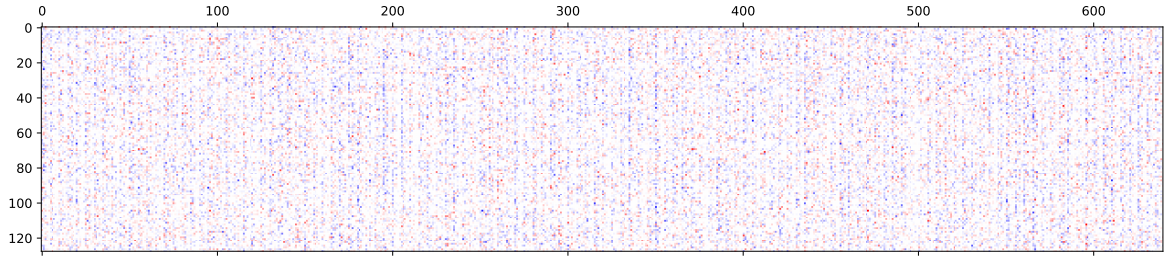
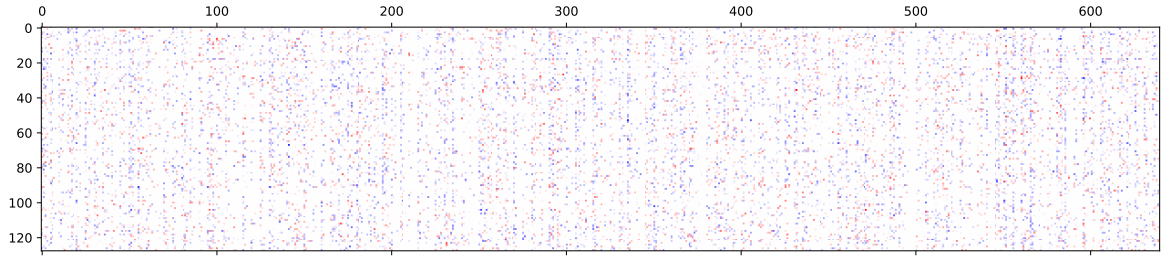
PENNI: Pruned Kernel Sharing for Efficient CNN Inference Appendix

1. Visualization of coefficient matrix

The coefficient matrix is shown in Fig.1. By conducting retraining, pruning and fine-tuning process, both weight and structured redundancies are explored.



(a) Decomposed



(b) Pruned

Figure 1. The visualization of the coefficient matrix. This is the third convolution layer of VGG16-CIFAR10 model with 128 input channels and 128 convolutional filters. The y-axis represents the output channels while the x-axis is the basis of each input channel.

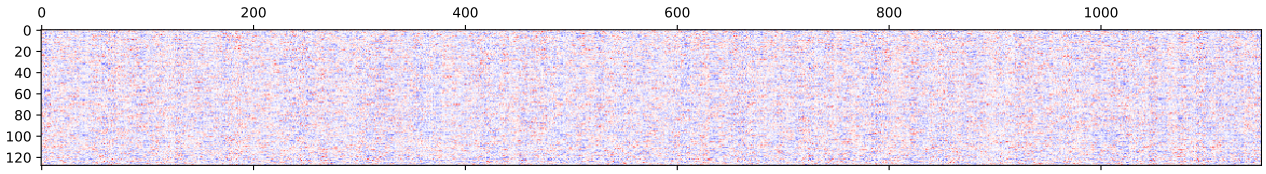
To compare with the baseline model, we reconstruct the weight matrix by multiplying the coefficient matrix with the basis. The reconstructed weight matrix is shown in Fig.2. Although the sparsity is also shown in the reconstructed weight, more computation can be saved by using the decomposed convolution.

2. Pruned Model Structure

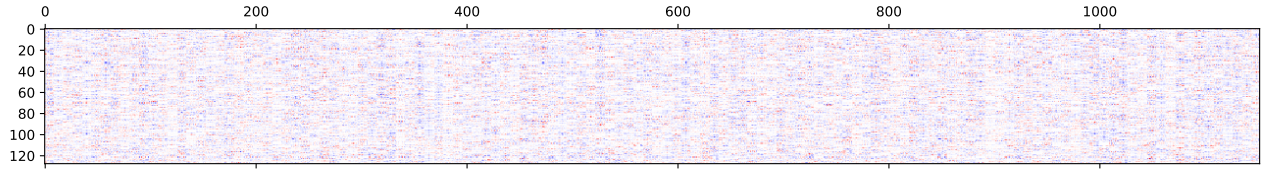
In this section, we list the detailed structure of the pruned model, including the number of basis kernels, the layer width before/after the shrinking process and the sparsity level of the coefficient matrix. We only consider the convolutional layers.

2.1. VGG16 on CIFAR10

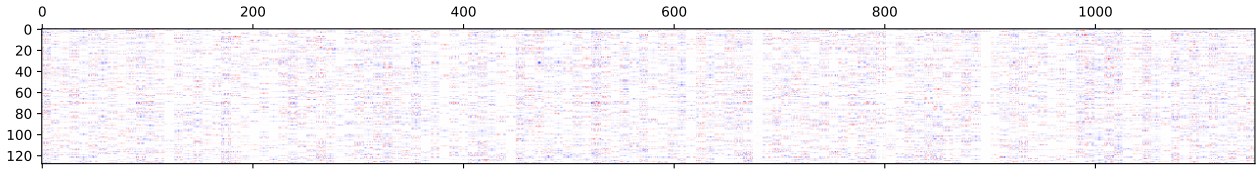
We list the detailed structure of the VGG16 before and after the shrinking process on Table.1. We only consider all the convolutional layers. The first index of the name represents the downsampling stage. The width represents the number of the filters (i.e., the number of output channels).



(a) Baseline



(b) Decomposed



(c) Pruned

Figure 2. The visualization of the reconstructed weight matrix comparing with the baseline model. This is the third convolution layer of VGG16-CIFAR10 model with 128 input channels and 128 convolutional filters. The y-axis represents the output channels.

Table 1. Detailed Structure of VGG16-CIFAR10

Name	Size	Before shrinking				After shrinking			
		Width	Basis	Coefficients (Non-zero/Total)	Sparsity/%	Width	basis	Coefficients (Non-zero/Total)	Sparsity/%
conv1_1	3	64	5	196/960	79.58	55	5	196/825	76.24
conv1_2	3	64	5	2313/20480	88.71	64	5	2238/17600	87.28
conv2_1	3	128	5	5862/40960	85.69	128	5	5862/40960	85.68
conv2_2	3	128	5	12052/81920	85.29	128	5	12052/81920	85.28
conv3_1	3	256	5	23169/163840	85.86	256	5	23169/163840	85.85
conv3_2	3	256	5	36870/327680	88.75	256	5	36870/327680	88.74
conv3_3	3	256	5	27719/327680	91.54	237	5	27716/303360	90.86
conv4_1	3	512	5	15688/65536	97.61	158	5	15665/187230	91.63
conv4_2	3	512	5	4534/1310720	99.65	62	5	4530/48980	90.75
conv4_3	3	512	5	2668/1310720	99.79	48	5	2666/14880	82.08
conv5_1	3	512	5	1318/1310720	99.89	36	5	1315/8640	84.78
conv5_2	3	512	5	874/1310720	99.93	26	5	874/4680	81.32
conv5_3	3	512	5	2554/1310720	99.80	486	5	2554/63180	95.95
Total				135817/8172480	98.34			135707/1263775	89.26

2.2. ResNet50 on ImageNet

We list the detailed structure of ResNet50-R1 after shrinking process in Table.2. We only consider all the convolutional layers including the 1×1 convolution. 'L' represents the downsampling stages while the 'B' stands for the residual blocks. This table reveals that, although for some layer the reduction of the layer width is not significant, the model can still benefit from pruning basis kernels.

Table 2: Detailed Structure of ResNet50 after the shrinking process.

Name	Size	width	# of basis	coefficients (non-zero/total)	sparsity/%
conv1	7	59	5	128/885	85.53
L1B1.conv1	1	22	-	500/1289	61.48
L1B1.conv2	3	29	4	125/2552	95.1
L1B1.conv3	1	256	-	2223/7424	70.06
L1B1.downsample	1	256	-	2589/15104	82.86
L1B2.conv1	1	32	-	2636/8192	67.82
L1B2.conv2	3	34	5	349/5440	93.58
L1B2.con3	1	256	-	2301/8704	73.56
L1B3.conv1	1	31	-	2978/7936	62.47
L1B3.conv2	3	63	5	807/9756	91.74
L1B3.conv3	1	256	-	3395/16128	78.95
L2B1.conv1	1	88	-	5973/22528	73.49
L2B1.conv2	3	115	3	581/30360	98.09
L2B1.conv3	1	512	-	11939/58880	79.72
L2B1.downsample	1	512	-	23134/131072	82.35
L2B2.conv1	1	20	-	3778/10240	63.11
L2B2.conv2	3	76	4	334/6080	94.51
L2B2.conv3	1	512	-	8154/38912	79.05
L2B3.conv1	1	66	-	9898/33792	70.71
L2B3.conv2	3	98	4	497/25872	98.08
L2B3.conv3	1	512	-	10702/50176	78.67
L2B4.conv1	1	78	-	12508/39936	68.68
L2B4.conv2	3	95	4	1006/29640	96.61
L2B4.conv3	1	512	-	12680/48640	73.93
L3B1.conv1	1	240	-	29283/122880	76.17
L3B1.conv2	3	229	2	712/100920	99.35
L3B1.conv3	1	1024	-	57763/234496	75.37
L3B1.downsample	1	1024	-	117775/524288	77.54
L3B2.conv1	1	103	-	29283/122880	76.17
L3B2.conv2	3	182	5	866/93730	99.08
L3B2.conv3	1	1024	-	57763/234496	75.37
L3B3.conv1	1	94	-	35956/96256	62.65
L3B3.conv2	3	180	5	1222/84600	98.56
L3B3.conv3	1	1024	-	53268/184320	71.1
L3B4.conv1	1	135	-	49843/138240	63.94
L3B4.conv2	3	193	5	925/130275	99.29
L3B4.conv3	1	1024	-	58240/197632	70.53
L3B5.conv1	1	155	-	56279/158720	64.54
L3B5.conv2	3	197	4	1138/122140	99.07
L3B5.conv3	1	1024	-	60387/201728	70.07
L3B6.conv1	1	206	-	65496/210944	68.95
L3B6.conv2	3	217	4	1156/178808	99.35
L3B6.conv3	1	2024	-	65842/222208	70.37
L4B1.conv1	1	495	-	148203/506880	70.76
L4B1.conv2	3	484	1	546/239580	99.77
L4B1.conv3	1	2048	-	279619/991232	91.79
L4B1.downsample	1	2048	-	580155/2097152	72.34
L4B2.conv1	1	411	-	276731/841728	67.12

Table 2 continued from previous page

Nam	Size	width	# of basis	coefficients (non-zero/total)	sparsity/%
L4B2.conv2	3	445	4	1334/731580	99.82
L4B2.conv3	1	2048	-	275688/911360	69.75
L4B3.conv1	1	501	-	307600/1026048	70.02
L4B3.conv2	3	484	3	1076/727452	99.85
L4B3.conv3	1	2048	-	245290/991232	75.25
total				2.98M/12.98M	77.0

3. Learning Curves

We show the learning curve of both the retraining and fine-tuning phase of the ResNet-50 model on ImageNet dataset. The curve here belongs to the 'R4' setting, which is $d = 6$ and regularization strength $\lambda = 5e - 5$

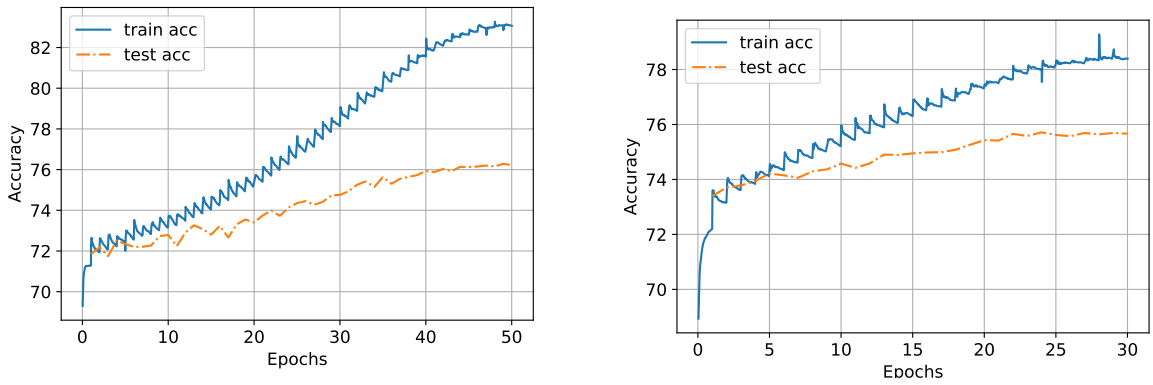


Figure 3. The learning curve of the *(left)* retraining phase and *(right)* fine-tuning phase. The model is ResNet-50 with parameter setting R4.